Today

- HW 3 due April 1
- Project due April 15
- In the News: more on Vitamin D
- nonparametric and semi-parametric regression

Recap: Regression splines

- regression splines are 'hand-crafted': `bs(x, df = ?)`
- once crafted, model is parametric
- $y_i = f(x_i) + \epsilon_i = \sum_{m=1}^{M} \beta_m h_m(x_i) + \epsilon_i$ `df` = M
- degrees of freedom controls how many knots are placed
- equivalent to how many parameters are fit
- ELM, §11.2 constructs linear splines 'by hand' Figure 11.7
- a different approach is smoothing splines – put a LOT of knots, and regularize

Smoothing splines

ELM §11.2, SM §10.7.2

- $y_i = f(x_i) + \epsilon_i, \quad i = 1, \ldots, n$
- choose $f(\cdot)$ to solve

$$\min_f \sum_{i=1}^{n} \frac{\{y - f(x_i)\}^2}{2\sigma^2} - \frac{\lambda}{2\sigma^2} \int_a^b \{f''(t)\}^2 dt, \quad , \lambda > 0$$

- solution is a cubic spline, with knots at each observed x_i value

see SM Figure 10.18 for a non-regularized solution

- has an explicit, finite dimensional solution
- $\hat{f} = \{\hat{f}(x_1), \ldots, \hat{f}(x_n)\} = (I + \lambda K)^{-1} y = \mathcal{S}_\lambda y$ as with local polynomials
- K is a symmetric $n \times n$ matrix of rank $n - 2$
- minimize over all differentiable functions with an absolutely continuous first derivative

Multidimensional splines

- so far we are considering just 1 explanatory variable

$$y_i = f(x_i) + \epsilon_i$$

- thin plate splines smooth two or more inputs simultaneously, but computational difficulty increases rapidly

- kernel smoothing can handle bivariate inputs as well

ELM §11.7

- additive models: $y_i = f_1(x_{1i}) + f_2(x_{2i}) + \cdots + f_p(x_{pi}) + \epsilon_i$

ELM Ch. 12

- binary response: $\text{logit}(p_i) = f_1(x_{1i}) + f_2(x_{2i}) + \cdots + f_p(x_{pi})$

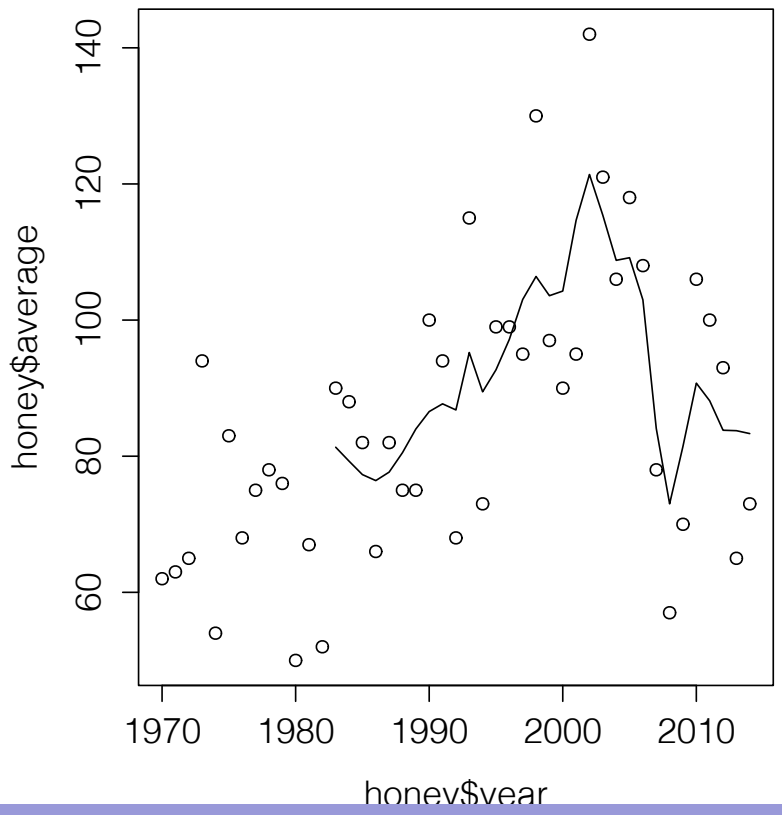
- generalized additive models:
  $g\{\mathsf{E}(y_i)\} = f_1(x_{1i}) + f_2(x_{2i}) + \cdots + f_p(x_{pi})$

Wavelets

ELM §11.4

- basis functions with very 'local' support
- useful for capturing regions of high variation
- useful for data compression – transmit only local components

Fig 11.11



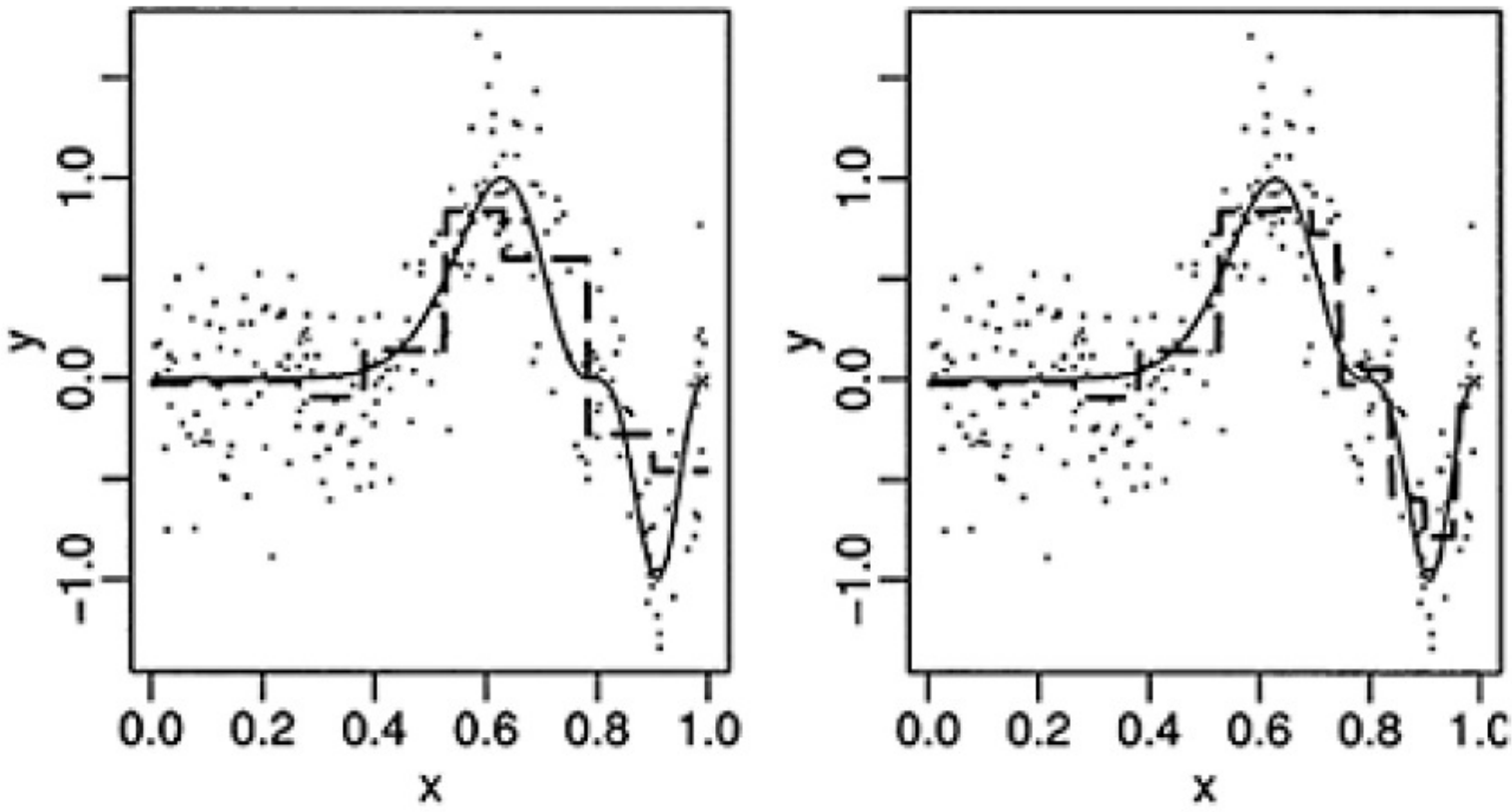
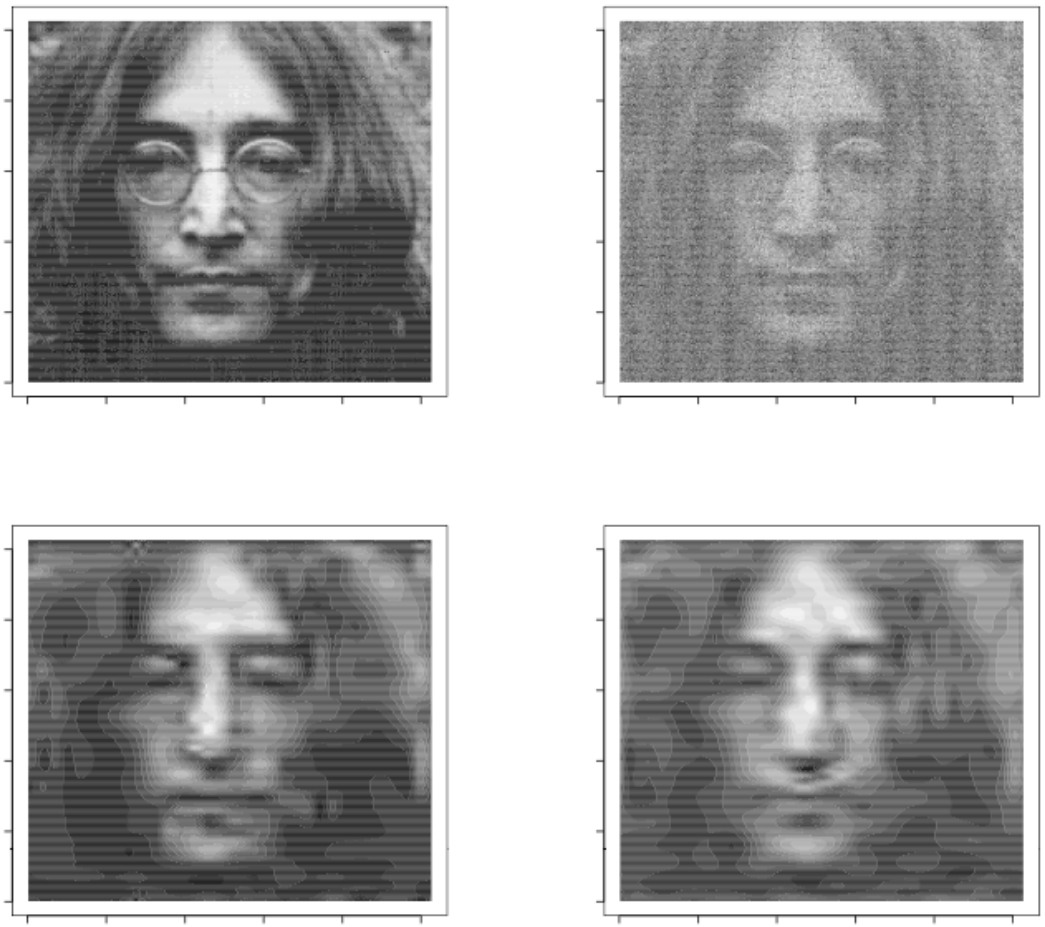


Figure 11.11 *Thresholding and inverting the transform. In the left panel all level-four and above coefficients are zeroed. In the right, the coefficients are thresholded using the default method. The true function is shown as a solid line and the estimate as a dashed line.*

... wavelets



Vidaković and Mueller, ”Wavelets for kids (Part I)” 1994.

Which smoothing method?

ELM §11.6

- basis functions: natural splines, Fourier, wavelet bases
- regularization via cubic smoothing splines
- kernel smoothers: locally constant/linear/polynomial
- adaptive bandwidth, running medians, running M-estimates
- with very little noise, a small amount of local smoothing (e.g. nearest neighbours)
- with moderate amounts of noise, kernel and spline methods are effective
- with large amounts of noise, parametric methods are more attractive
- "It is not reasonable to claim that any one smoother is better than the rest"
  - `loess` is robust to outliers, and provides smooth fits
  - spline smoothers are more efficient, but potentially sensitive to outliers
  - kernel smoothers are very sensitive to bandwidth

Shrinkage Methods

Elements of Statistical Learning, Hastie et al. §3.4

- Ridge regression
-

$$\begin{aligned} \hat{\beta}_{LS} &= (X^T X)^{-1} X^T y \\ \hat{\beta}_{ridge} &= (X^T X + \lambda I)^{-1} X^T y \end{aligned}$$

- can show that $\hat{\beta}_{ridge}$ satisfies

$$\min_{\beta} \left(\Sigma \{ y_i - \beta_0 - \Sigma_{j=1}^{p} x_{ij} \beta_j \}^2 + \lambda \Sigma_{j=1}^{p} \beta_j^2 \right)$$

$$\min_{\beta} \Sigma \{ y_i - \beta_0 - \Sigma_{j=1}^{p} x_{ij} \beta_j \}^2 \quad \text{s.t. } \Sigma \beta_j^2 \leq t$$

- Assume x_j's are centered and put these in matrix X (with no column of 1's:

$$\min_{\beta} (y - X\beta)^T (y - X\beta) \qquad \text{s.t. } ||\beta||^2 \leq t$$

... ridge regression

- $$\min_{\beta}\{(y - X\beta)^T(y - X\beta) + \lambda||\beta||^2\}$$
- λ is a tuning parameter: $\lambda = 0$ gives $\hat{\beta}_{LS}$, $\lambda \to \infty$
- in R the `library MASS` `library(MASS)` has a ridge regression version of `lm` called `lm.ridge`
- if columns of X are nearly linearly dependent (multicollinearity), $\hat{\beta}$'s for these columns should be shrunk towards 0.
- essential that the predictors are all scaled to the same units
- this is difficult for interpretation of the coefficients
- $$df(\lambda) = \text{tr}[X(X^TX + \lambda I)^{-1}X^T] = \Sigma_{j=1}^{p} \frac{d_j^2}{d_j^2 + \lambda}$$

effective number of parameters

Lasso

- $$\min_{\beta} \left(\Sigma \{ y_i - \beta_0 - \Sigma_{j=1}^{p} x_{ij} \beta_j \}^2 + \lambda \Sigma_{j=1}^{p} |\beta_j| \right)$$
- $$\min_{\beta} \Sigma \{ y_i - \beta_0 - \Sigma_{j=1}^{p} x_{ij} \beta_j \}^2 \quad \text{s.t. } \Sigma |\beta_j| \leq t$$
- quadratic programming problem
- $\hat{\beta}^{lasso}$ is nonlinear function of y
- Tibshirani (1996), JRSS B and (2011), JRSS B
- `http://http://www-stat.stanford.edu/~tibs/lasso.html`

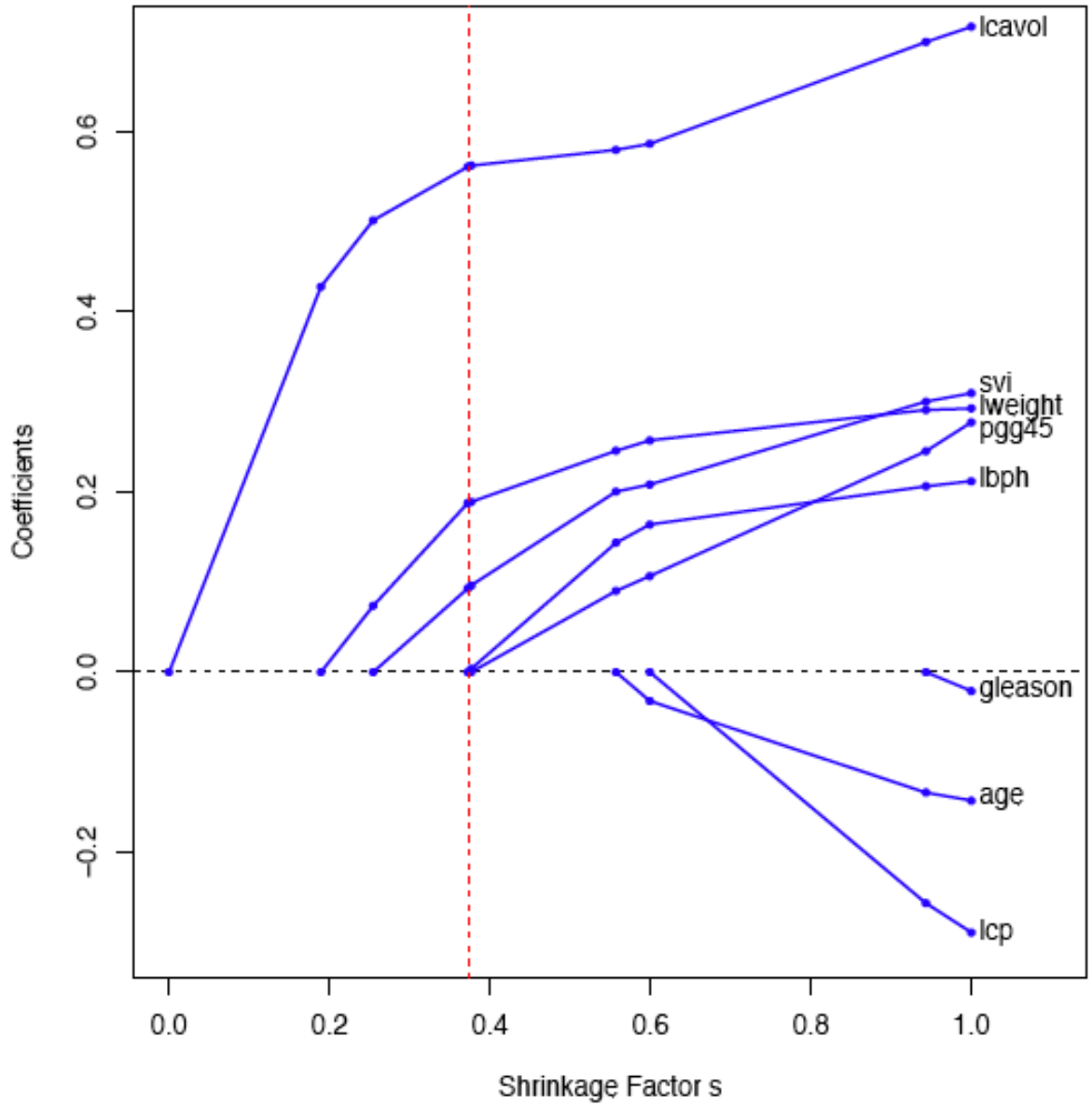


FIGURE 3.10. *Profiles of lasso coefficients, as the tuning parameter t is varied. Coefficients are plotted versus $s = t/\sum_1^p |\hat{\beta}_j|$. A vertical line is drawn at $s = 0.36$, the value chosen by cross-validation. Compare Figure 3.8 on page 65; the lasso*

... shrinkage

- ridge regression gives "proportional shrinkage"
- subset selection gives "hard thresholding" (some $\beta_j \to 0$)
- lasso gives "soft thresholding": blend of shrinkage and zeroing

- elastic net combines lasso and ridge regression

$$\min_{\beta} \left(\sum \{y_i - \beta_0 - \Sigma_{j=1}^{p} x_{ij}\beta_j\}^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2 \right)$$

- implemented in `R` in `library(glmnet)`

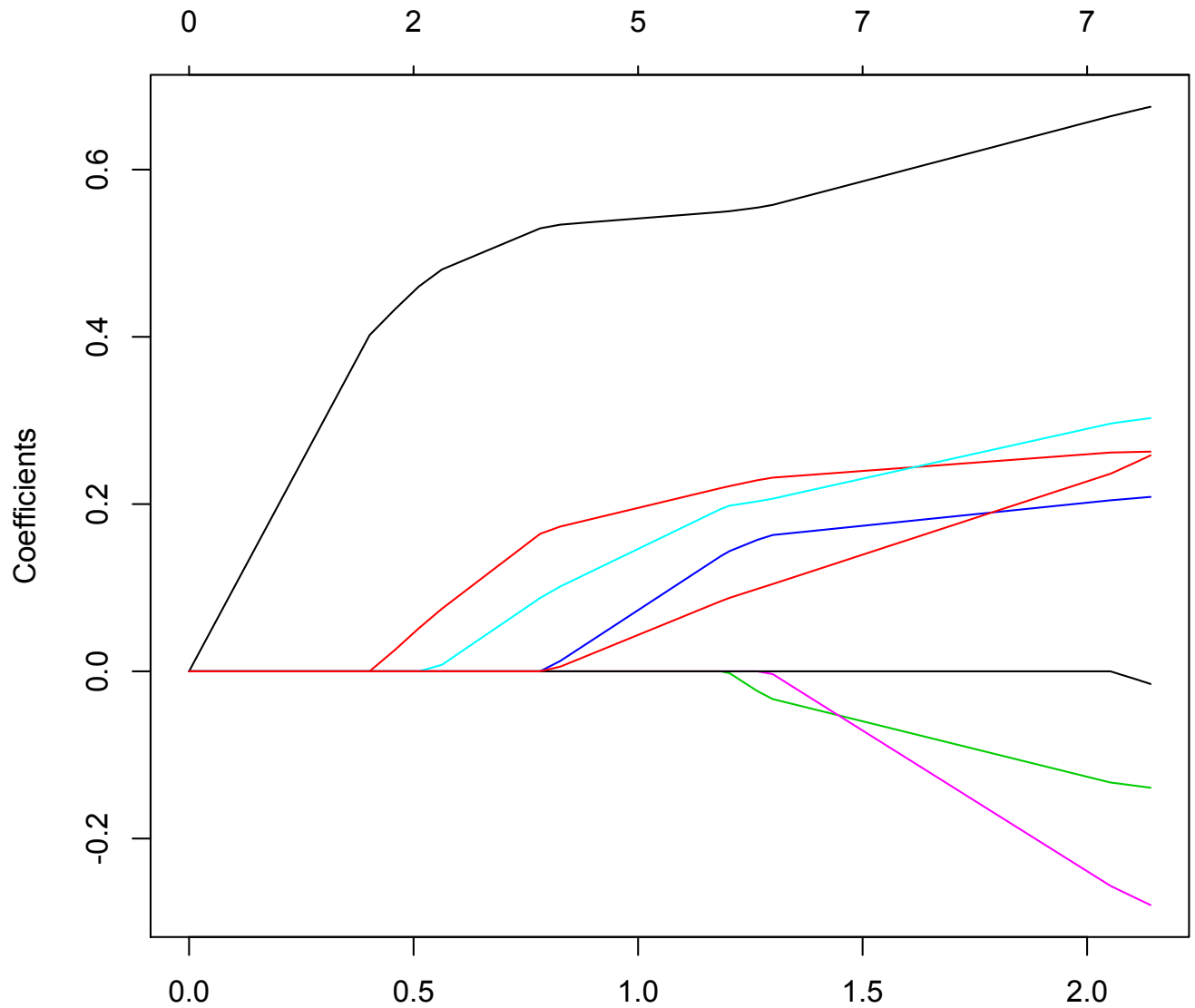
- estimates of coefficients are biased (but may have small mean-squared error)
- Lasso is now used as a variable selection method
- improvements in algorithms allow fast computation even for $p > n$

Prostate data

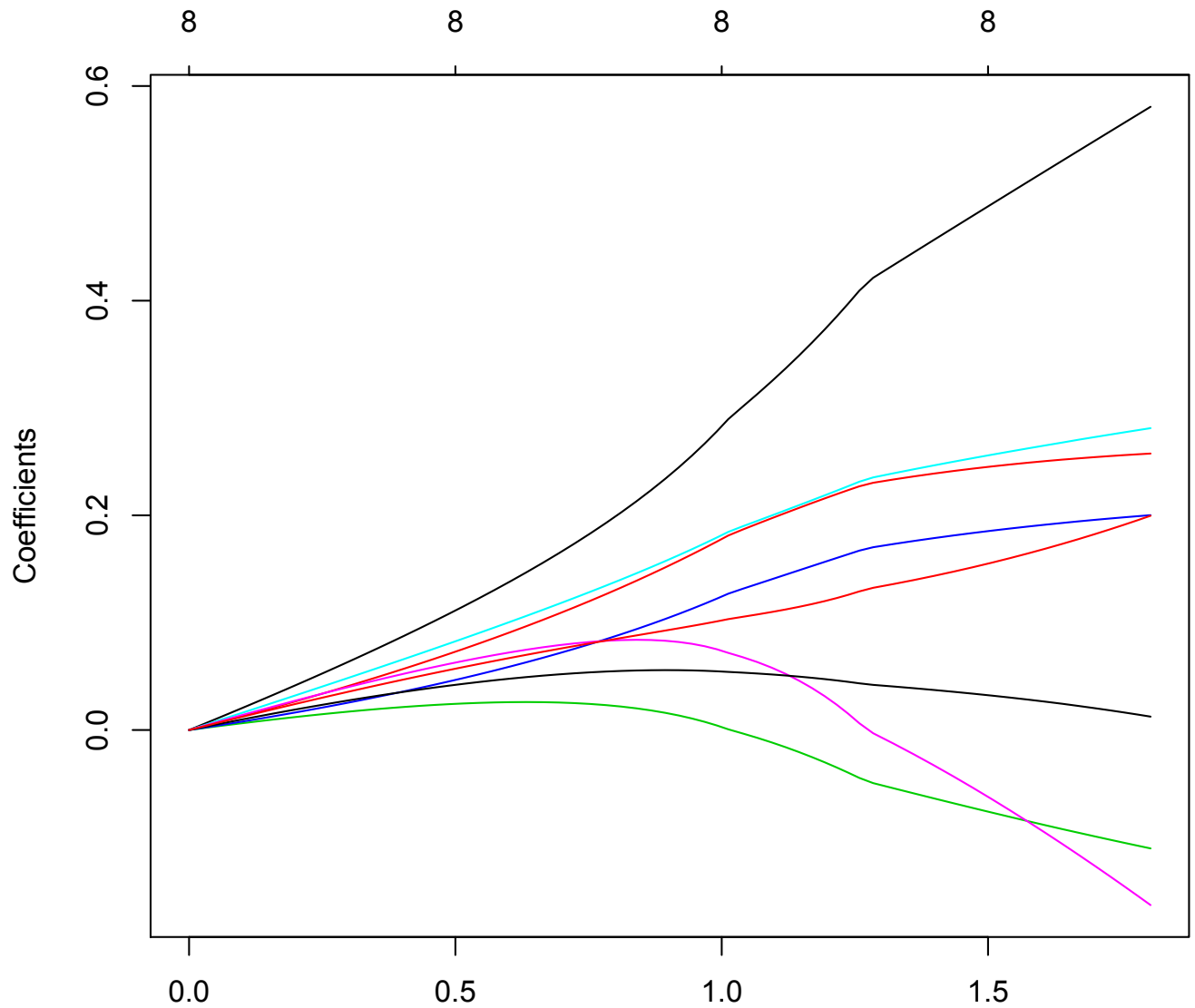
Ch.3, HTF

```
> prostate <- read.csv(file="prostate.data",sep="\t")
## data is at http://statweb.stanford.edu/~tibs/ElemStatLearn/datasets/prostate.data

> head(prostate)
      lcavol  lweight age      lbph svi       lcp gleason pgg45
1 -0.5798185 2.769459  50 -1.386294   0 -1.386294       6     0
2 -0.9942523 3.319626  58 -1.386294   0 -1.386294       6     0
3 -0.5108256 2.691243  74 -1.386294   0 -1.386294       7    20
4 -1.2039728 3.282789  58 -1.386294   0 -1.386294       6     0
5  0.7514161 3.432373  62 -1.386294   0 -1.386294       6     0
6 -1.0498221 3.228826  50 -1.386294   0 -1.386294       6     0
        lpsa train
1 -0.4307829  TRUE
2 -0.1625189  TRUE
3 -0.1625189  TRUE
4 -0.1625189  TRUE
5  0.3715636  TRUE
6  0.7654678  TRUE

> xp <- scale(prostate[,1:8])
> y <- prostate[,9]
> train <- prostate[,10]
## standardize data; y is the response (log psa); extract training data
##
> library(glmnet)
> pr.lasso <- glmnet(xp[train,],y[train]) #Lasso
> pr.lasso2 <- glmnet(xp[train,],y[train], alpha=0) # ridge
> plot(pr.lasso); plot(pr.lasso2)
```

... prostate data



... prostate data

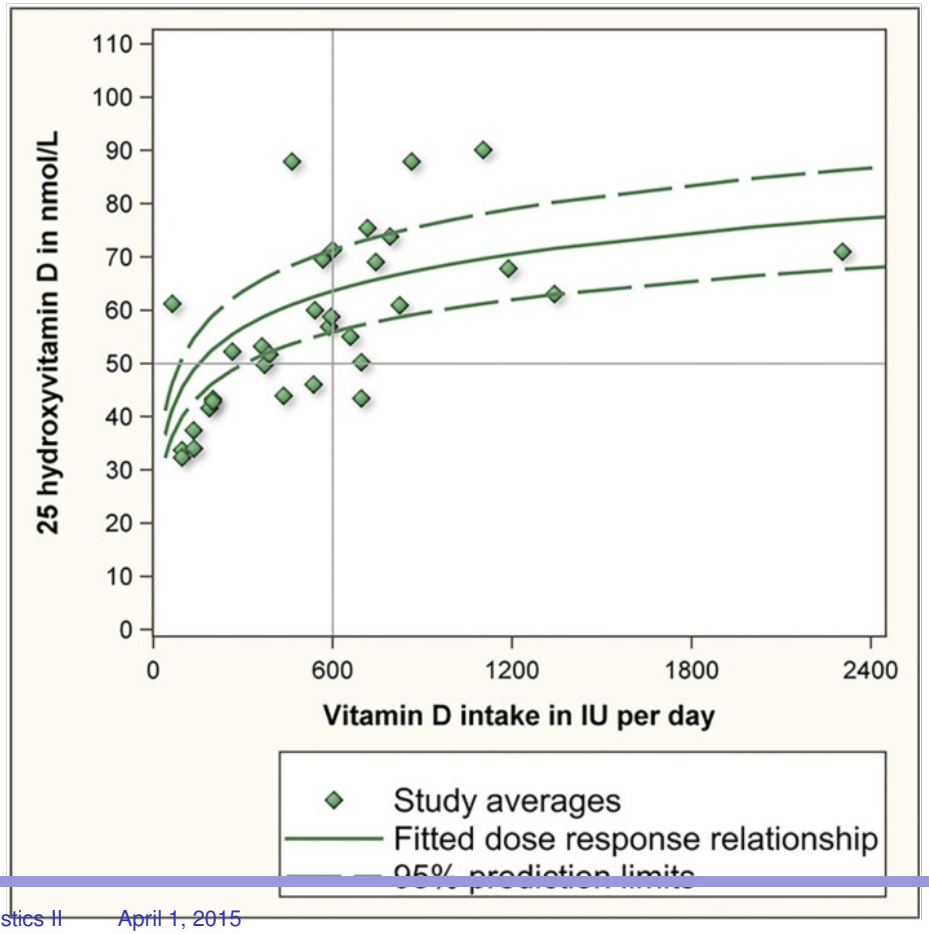


Vitamin D again

PMC full text: Nutrients. 2014 Oct; 6(10): 4472–4475.
Published online 2014 Oct 20. doi: 10.3390/nu6104472

<< Prev Figure 1 Next >>

Figure 1



The vitamin D dilemma: How much should we be taking?

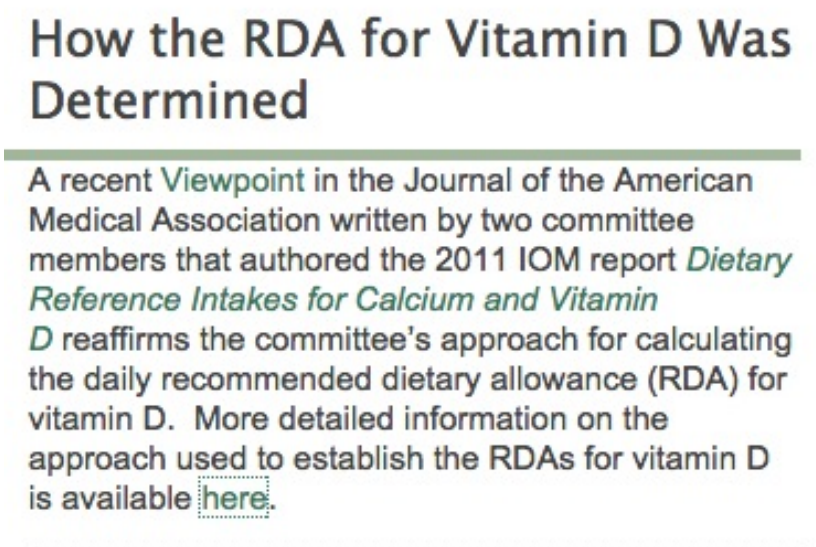
LESLIE BECK
Special to The Globe and Mail

Published Sunday, Mar. 29 2015, 12:00 PM EDT
Last updated Sunday, Mar. 29 2015, 12:00 PM EDT

True, it was a painful winter. We were stuck in a cocoon of cold (some of us

... vitamin D

Institute of Medicine



Announcement

How the RDA for Vitamin D Was Determined

The Recommended Dietary Allowance (RDA), by definition, meets the requirements of 97.5 percent of the population. It is set from an Estimated Average Requirement (EAR) that represents an intake amount that will meet the needs of about 50 percent of the population. The RDA represents an intake amount that is 2 standard deviations above the EAR. What this means for vitamin D is that for most of the population, their requirements are met by an intake level that will achieve a serum level of the biomarker of vitamin D status, 25(OH)D of 50 nmol/L (equal to 20 ng/mL), determined to be equivalent to an RDA level of vitamin D intake (based on essentially all vitamin D from diet and minimal sunlight exposure), as illustrated in Figure 1 below.

Figure 1

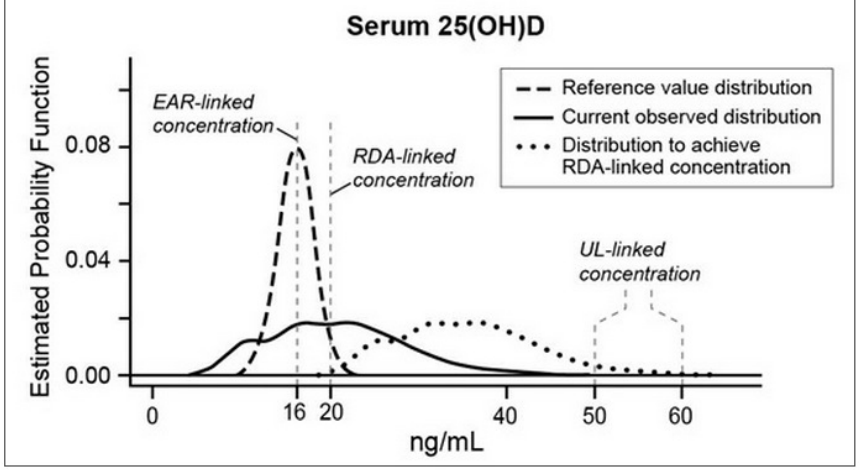


FIGURE 1: The Reference Value distribution of serum 25(OH)D concentrations (▬ ▬) displays the values linked with the Estimated Average Requirement (EAR) that meets the needs of 50% of the generally healthy populations, the Recommended Dietary Allowance (RDA) that meets the needs of 97.5% of the population and the Tolerable Upper Level (UL) above which the risk of adverse effects increases. This Reference Value distribution was derived by using the mean (95th percentile) specified by the Institute of Medicine with calculated SD = 2.0 ng/ml on the basis of normality; the estimated probability function (y-axis) indicates the frequency of each concentration in the sample. Also displayed is the current observed distribution of serum 25(OH)D concentrations (▬▬) for adults aged 19-70 y in NHANES 2005-2006 (n=3871). For comparison, the shift in the distribution to achieve serum 25(OH)D concentrations above the RDA-linked value of 20 ng/ml in 97.%% of the population (••••••) is also shown. Reproduced from Brannon PM, Mayne ST, Murphy SP, Taylor CL. Vitamin D Supplementation in African Americans: Dose-response. American Journal of Clinical Nutrition 2014;100:982-4 with permission from *American Journal of Clinical Nutrition*.

In its analysis, the IOM committee used a mixed-model approach to estimate the dose-relationship between total dietary intake of vitamin D to achieve the desired serum level of 25(OH)D. In its methodology to determine DRIs for vitamin D, the IOM committee used an estimated dose-response to inform their judgment of the intake needed to achieve serum 25(OH)D levels in the desired range. Because of limitations in the available data, the IOM committee made the judgment to overestimate the dose of dietary vitamin D needed; 600 IU daily would meet the RDA based on a serum 25(OH)D level of 50 nmol/L derived from the estimated dose-response relationship.

In this approach, the Estimated Average Requirement was determined to be equivalent to an intake that produced a serum level of 40 nmol/L (based on essentially all vitamin D from diet and minimal sunlight exposure). Thus, because this level represents the RDA for dietary vitamin D, the 2 standard deviations has, in effect, already been added and therefore no additional adjustment is required or appropriate. Using the distribution to 97.5 percent to assure a low prevalence of inadequacy among the healthy population groups, the RDA value was based on an achieved level of 50 nmol/L (equal to 20 ng/mL). To be clear, the goal is not, and should not be, to assure that 97.5% of the population exceeds the serum value linked to the RDA. Doing so would shift the distribution to a higher level that is associated with increased risk for adverse effects as illustrated in Figure 1 as Distribution to achieve RDA-linked concentration.

JAMA The Journal of the American Medical Association

Viewpoint | February 19, 2015

Vitamin D Research and Clinical Practice

At a Crossroads FREE ONLINE FIRST

JoAnn E. Manson, MD, DrPH[1]; Shari S. Bassuk, ScD[1]

[+] Author Affiliations

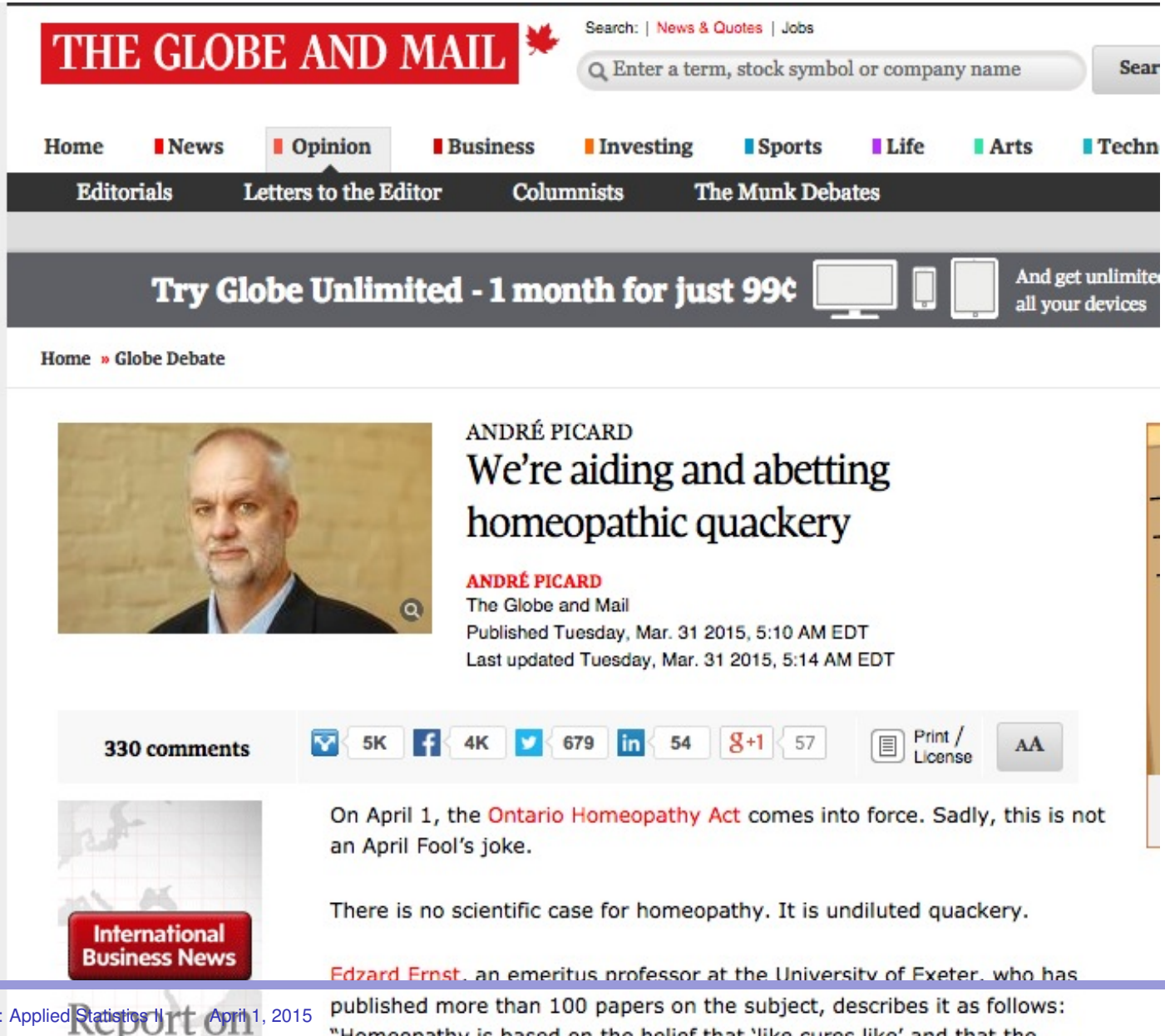
JAMA. Published online February 19, 2015. doi:10.1001/jama.2015.1353

Long recognized as important for bone health, vitamin D has attracted recent interest for its possible nonskeletal benefits. Many primary care clinicians now include blood tests to measure vitamin D concentrations as part of routine laboratory work[1] and recommend vitamin D supplements, often at high doses, to their patients for the possible prevention of cancer, cardiovascular disease (CVD), diabetes, autoimmune disorders, cognitive decline, and other conditions. Thus, screening rates and sales of vitamin D supplements have increased substantially in recent years.[1,2]

However, clinical enthusiasm for supplemental vitamin D has outpaced available evidence on its effectiveness and threatens to jeopardize the ability of researchers to conduct randomized trials in "usual-risk" populations. Based on its recent systematic reviews of the literature, the US Preventive Services Task Force (USPSTF) concluded that data are insufficient to recommend vitamin D screening in routine clinical practice[1] or to assess the effectiveness and overall balance of benefits and risks of supplemental vitamin D taken for the primary prevention of cancer and CVD.[3] In an earlier review, the Institute of Medicine (IOM) reached the same conclusion—namely, whether supplemental vitamin D lowers risk of nonskeletal health outcomes, and what dose might be required to do so, is uncertain.[4]

Homeopathy again

Link



THE GLOBE AND MAIL

Search: | News & Quotes | Jobs

Enter a term, stock symbol or company name

Sear

Home | News | Opinion | Business | Investing | Sports | Life | Arts | Techn

Editorials | Letters to the Editor | Columnists | The Munk Debates

Try Globe Unlimited - 1 month for just 99¢

And get unlimite
all your devices

Home » Globe Debate

ANDRÉ PICARD

We're aiding and abetting homeopathic quackery

ANDRÉ PICARD
The Globe and Mail
Published Tuesday, Mar. 31 2015, 5:10 AM EDT
Last updated Tuesday, Mar. 31 2015, 5:14 AM EDT

330 comments | 5K | 4K | 679 | 54 | g+1 57 | Print / License | AA

International Business News

Report on

On April 1, the Ontario Homeopathy Act comes into force. Sadly, this is not an April Fool's joke.

There is no scientific case for homeopathy. It is undiluted quackery.

Edzard Ernst, an emeritus professor at the University of Exeter, who has published more than 100 papers on the subject, describes it as follows:

Australian Government
National Health and Medical Research Council

NHMRC Statement: Statement on Homeopathy

Based on the assessment of the evidence of effectiveness of homeopathy, NHMRC concludes that there are no health conditions for which there is reliable evidence that homeopathy is effective.

Homeopathy should not be used to treat health conditions that are chronic, serious, or could become serious. People who choose homeopathy may put their health at risk if they reject or delay treatments for which there is good evidence for safety and effectiveness. People who are considering whether to use homeopathy should first get advice from a registered health practitioner.* Those who use homeopathy should tell their health practitioner and should keep taking any prescribed treatments.

The National Health and Medical Research Council expects that the Australian public will be offered treatments and therapies based on the best available evidence.

Homeopathy

Homeopathy is a type of complementary and alternative medicine. It is based on two main ideas: that substances that may cause illness or symptoms in a healthy person can, in very small doses, treat those symptoms in a person who is unwell; and that molecules in highly diluted substances retain a memory of the original substance.

NHMRC's methods

The National Health and Medical Research Council (NHMRC) undertook an assessment of the evidence of the effectiveness of homeopathy for treating health conditions. This assessment was based on:

- an overview of published systematic reviews by an independent contractor;
- an independent evaluation of information provided by homeopathy interest groups and the public; and
- consideration of clinical practice guidelines and government reports on homeopathy published in other countries.

The assessment of the evidence used standardised, accepted methods for assessing the quality and reliability of evidence for whether or not a therapy is effective for treating health conditions.